BMJ

RESEARCH

Multiple component patient safety intervention in English hospitals: controlled evaluation of second phase

Amirta Benning, programme manager,[1] Mary Dixon-Woods, professor of medical sociology,[2] Ugochi Nwulu, senior research associate/coordinator,[3] Maisoon Ghaleb, lecturer in pharmacy practice/patient safety,[4,5] Jeremy Dawson, research fellow,[6] Nick Barber, professor of the practice of pharmacy,[4] Bryony Dean Franklin, professor of medication safety and director, centre for medication safety and service quality,[4,7] Alan Girling, senior research fellow,[1] Karla Hemming, senior research fellow,[1] Martin Carmalt, consultant physician,[8] Gavin Rudge, data scientist,[1] Thirumalai Naicker, honorary research associate,[1] Amit Kotecha, registrar,[8] M Clare Derrington, independent contractor, public health physician,[9] Richard Lilford, professor of clinical epidemiology[1]

[1]School of Health and Population Sciences, University of Birmingham, Edgbaston, West Midlands B15 2TT, UK
[2]Department of Health Sciences, University of Leicester, Leicester LE1 7RH
[3]Clinical Investigation Unit, University Hospitals Birmingham NHS Foundation Trust, Queen Elizabeth Hospital, Birmingham B15 2TH
[4]Department of Practice and Policy, School of Pharmacy, University of London, London WC1N 1AX
[5]School of Pharmacy, University of Hertfordshire, Hatfield AL10 9AB
[6]Work and Organisational Psychology Group, Aston Business School, Aston University, Birmingham B4 7ET
[7]Imperial College Healthcare N HS Trust, St Mary's Hospital, London W2 1NY
[8]Royal Orthopaedic Hospital, Northfield, Birmingham B31 2AP
[9]45 Weston Road, Edith Weston, Rutland LE15 8HQ

Correspondence to: R J Lilford
r.j.lilford@bham.ac.uk

Cite this as: *BMJ* 2011;342:d199
doi:10.1136/bmj.d199

ABSTRACT

Objective To independently evaluate the impact of the second phase of the Health Foundation's Safer Patients Initiative (SPI2) on a range of patient safety measures.

Design A controlled before and after design. Five substudies: survey of staff attitudes; review of case notes from high risk (respiratory) patients in medical wards; review of case notes from surgical patients; indirect evaluation of hand hygiene by measuring hospital use of handwashing materials; measurement of outcomes (adverse events, mortality among high risk patients admitted to medical wards, patients' satisfaction, mortality in intensive care, rates of hospital acquired infection).

Setting NHS hospitals in England.

Participants Nine hospitals participating in SPI2 and nine matched control hospitals.

Intervention The SPI2 intervention was similar to the SPI1, with somewhat modified goals, a slightly longer intervention period, and a smaller budget per hospital.

Results One of the scores (organisational climate) showed a significant (P=0.009) difference in rate of change over time, which favoured the control hospitals, though the difference was only 0.07 points on a five point scale. Results of the explicit case note reviews of high risk medical patients showed that certain practices improved over time in both control and SPI2 hospitals (and none deteriorated), but there were no significant differences between control and SPI2 hospitals. Monitoring of vital signs improved across control and SPI2 sites. This temporal effect was significant for monitoring the respiratory rate at both the six hour (adjusted odds ratio 2.1, 99% confidence interval 1.0 to 4.3; P=0.010) and 12 hour (2.4, 1.1 to 5.0; P=0.002) periods after admission. There was no significant effect of SPI for any of the measures of vital signs. Use of a recommended system for scoring the severity of pneumonia improved from 1.9% (1/52) to 21.4% (12/56) of control and from 2.0% (1/50) to 41.7% (25/60) of SPI2 patients. This temporal change was significant (7.3, 1.4 to 37.7; P=0.002), but the difference in difference was not significant (2.1, 0.4 to 11.1; P=0.236). There were no notable or significant changes in the pattern of prescribing errors, either over time or between control and SPI2 hospitals. Two items of medical history taking (exercise tolerance and occupation) showed significant improvement over time, across both control and SPI2 hospitals, but no additional SPI2 effect. The holistic review showed no significant changes in error rates either over time or between control and SPI2 hospitals. The explicit case note review of perioperative care showed that adherence rates for two of the four perioperative standards targeted by SPI2 were already good at baseline, exceeding 94% for antibiotic prophylaxis and 98% for deep vein thrombosis prophylaxis. Intraoperative monitoring of temperature improved over time in both groups, but this was not significant (1.8, 0.4 to 7.6; P=0.279), and there were no additional effects of SPI2. A dramatic rise in consumption of soap and alcohol hand rub was similar in control and SPI2 hospitals (P=0.760 and P=0.889, respectively), as was the corresponding decrease in rates of *Clostridium difficile* and meticillin resistant *Staphylococcus aureus* infection (P=0.652 and P=0.693, respectively). Mortality rates of medical patients included in the case note reviews in control hospitals increased from 17.3% (42/243) to 21.4% (24/112), while in SPI2 hospitals they fell from 10.3% (24/233) to 6.1% (7/114) (P=0.043). Fewer than 8% of deaths were classed as avoidable; changes in proportions could not explain the divergence of overall death rates between control and SPI2 hospitals. There was no significant difference in the rate of change in mortality in intensive care. Patients' satisfaction improved in both control and SPI2 hospitals on all dimensions, but again there were no significant changes between the two groups of hospitals.

Conclusions Many aspects of care are already good or improving across the NHS in England, suggesting considerable improvements in quality across the board. These improvements are probably due to contemporaneous policy activities relating to patient safety, including those with features similar to the SPI, and the emergence of professional consensus on some clinical processes. This phenomenon might have attenuated the incremental effect of the SPI, making it difficult to detect. Alternatively, the full impact of the SPI might be observable only in the longer term. The conclusion of this study could have been different if concurrent controls had not been used.

INTRODUCTION

In the first phase of the Health Foundation's Safer Patients Initiative (SPI1) four hospitals in the United Kingdom took part in an organisational intervention to "transform organisational approaches to delivering safer care" designed and mentored by the Institute for Healthcare Improvement (IHI) and implemented in an 18 month period from the end of 2004.[1] We report our evaluation of SPI1 in a companion paper.[1] A second phase of the intervention, known as SPI2, was rolled out over 20 months from March 2007 to September 2008 inclusive. SPI2 included 20 UK hospitals (10 in England and 10 in the other countries of the UK) selected with a process similar to that used for SPI1. The intervention itself was modelled on that used for SPI1,[1] with some modifications (box 1). A full report including methodological and analytical detail is available.[2]

METHODS

Our methods were similar to those used for the evaluation of SPI1.[2] As with SPI1, the SPI2 evaluation used a series of linked substudies to address both generic features of systems that might be expected to improve if a general strengthening of organisational systems in relation to patient safety occurred, and specific process outcomes that were targeted specifically by SPI interventions.

Box 1: Differences between SPI1 and SPI2

- The hospitals were required to work with a partner organisation (a "buddy system") and encouraged to hold regular meetings between the lead implementation teams (10-12 people) from each site
- There was a longer period between dissemination of the preparatory materials (December 2006) and the first "kick-off" session where the various teams came together with the Institute for Healthcare Improvement to share experiences (March 2007)
- The financial package was smaller than in the case of SPI1; a mean of £270 000 (€314 000, $430 000) per site, rather than £775 000 (€900 000, $1.2m)
- The adverse event target was revised from a reduction of 50% to a reduction of 30%
- SPI2 sought a 15% reduction in mortality rates; this was not an explicit SPI1 aim
- The routine use of β blockers in the surgical "bundle" was removed as this clinical standard was contentious in the UK

Table 1 | Summary of substudies in evaluation of phase two of Safer Patients Initiative (SPI2)

Substudy and topic	Data source	Unit of analysis
Staff survey*		
Staff morale, culture, and opinion	NHS national staff survey	Staff member
Quality of care: acute medical care*		
In patients aged >65 with acute respiratory disease	Case note reviews (both explicit and implicit)	Patient
Quality of care: perioperative care*		
In patients with total hip replacement and open colectomy	Explicit case note review	Patient
Clinical process measures*		
Use of consumables for hand hygiene	National observation study of effectiveness of national "cleanyourhands" campaign	Hospital
Outcomes		
Adverse events in patients aged >65 with acute respiratory disease*	Holistic case note review	Patient
Hospital mortality in patients aged >65 with acute respiratory disease*	Case note review	Patient
Intensive care unit mortality†	Routine data from intensive care national audit and research centre	Hospital
Infection rates associated with healthcare †	Routine data from Health Protection Agency	Hospital
Patient satisfaction*	NHS patient surveys	Patient‡

*Data collected and analysed centrally
†Data collected by hospital staff, then analysed centrally.
‡In SPI1 unit of analysis was "hospital" as, in that case, we did not have individual patient data.

Framework for evaluation

Table 1 summarises the substudies; all made use of a controlled before and after design.[3] While no qualitative data were collected in SPI2, all of the quantitative studies undertaken in the SPI1 evaluation were replicated. Additional quantitative substudies were added to address SPI objectives not directly studied in our evaluation of SPI1.[1] These included review of surgical case notes to measure compliance with a set of evidence based standards for perioperative surgical care; examination of outcome data from intensive care units to provide evidence relevant to effectiveness of SPI2 methods to improve adherence to evidence based guidelines for critical care; assessment of consumption of soap and alcohol hand rub in hospital trusts, along with measures of rates of infection with *Clostridium difficile* and meticillin resistant *Staphylococcus aureus* (MSRA) to provide evidence on measures to reduce infections associated with healthcare; and audit by two independent reviewers of all deaths in our case note review.

Intervention and control sites

To take advantage of routinely collected data, we focused on SPI2 hospitals in England only. One SPI2 hospital declined to participate in the evaluation,

Box 2: Criteria for selection of control sites

- Only non-specialist acute hospitals in England were considered
- Control and SPI2 hospitals should have a similar directorate structure (as described in the NHS national staff survey)
- Hospitals should have the same "foundation" or "non-foundation" status (to gain foundation status a hospital must satisfy the government that it has the management capacity to warrant greater operational autonomy)
- Hospitals should be similarly located in either urban or rural settings
- Once these criteria were satisfied, the hospital with the most similar size (usually within 1000 staff) to the SPI2 hospital was selected as the control hospital

leaving nine available for study. Nine matched control sites were selected with the criteria in box 2. Table 2 gives details of control and SPI2 hospitals. Control hospitals were selected and approached in August 2007, seven months after the intervention had started in SPI2 hospitals (and after it had been completed in SPI1 hospitals).

Substudies

We carried out five substudies (see table 1).

Staff surveys

Staff morale, attitudes, and the organisational climate might be expected to change in response to SPI2. All nine SPI2 study sites and nine control sites were included in both the 2006 and 2008 National Staff Surveys, conducted between October and December in each of these years. Methods were the same as those used for the evaluation of SPI1,[1] with a sample of 850 staff members per site, except that a further two new relevant survey questions, not available for the SPI1 evaluation, were included.

Table 2 | SPI2 and matched control hospitals in phase two of Safer Patients Initiative (SPI2)

Hospital No	Beds (hospital, current)	Area*	Teaching status†
SPI2 hospitals			
1	411	Rural	Affiliated
2	455	Urban	Nil
3	620	Urban/rural	Nil
4	634	Urban	Nil
5	688	Urban	Teaching hospital
6	804	Urban	Teaching hospital
7	668	Urban	Teaching hospital
8	523	Urban	Teaching hospital
9	566	Urban	Affiliated
Matched control hospitals			
1	475	Rural	Nil
2	511	Urban	Nil
3	618	Urban	Teaching hospital
4	723	Urban/rural	Nil
5	447	Urban/rural	Affiliated
6	789	Urban	Affiliated
7	988	Urban	Affiliated
8	532	Urban/rural	Nil
9	1036	Urban	Affiliated

*Based on visual inspection of population density map.
†According to hospital website.

Error rates/quality of acute medical care

We sought to assess improvement in error rates and quality of care. We selected patients aged over 65 with acute respiratory disease admitted to acute medical wards as the focus for study for the same reasons as in the SPI1 evaluation (high risk, error prone population). The areas of review included both those specifically targeted by SPI2, such as vigilance in monitoring sick patients and prescription error, and those that might be expected to improve if an overall shift in organisational systems and culture related to patient safety occurred, such as adherence to various tenets of evidence based care. Case notes were processed and audited with the same procedures and criteria used for the SPI1 evaluation.

Case notes were collected over three epochs (time periods). Observations before implementation of SPI2 were spread over two epochs (epoch 1: October 2003 to March 2004; epoch 2: October 2006 to March 2007). Epoch 3 (October 2008 to March 2009) was after the intervention. As described for SPI1, case notes were reviewed by MG, with one in 10 re-reviewed by BDF.

Using review against explicit criteria, we aimed to analyse 15 sets of case notes from each control and SPI2 hospital per epoch (810 in total). This provided 80% power to detect a 13 percentage point change in staff compliance with a standard where baseline compliance was 70%. In each month, from each epoch, we selected from each hospital the case notes from the first two or three patients who fulfilled the eligibility criteria. As before, case notes were not examined in series, allowing us to detect learning/fatigue effects among reviewers.

In addition to the criterion based explicit review, a specialist in general medicine (MC) with experience in reviewing case notes (the principal reviewer) evaluated each set holistically for evidence of errors and adverse events. To measure reliability, an experienced trainee in respiratory medicine (TN, the reliability reviewer) independently re-evaluated a subset of notes (n=74). Errors were analysed and categorised as in our SPI1 evaluation.[1] In addition, each death was re-analysed by a second reviewer (MCD), medically trained in anaesthetics and public health ("blinded" to epoch and group), who had experience as a reviewer for the National Confidential Enquiry into Perioperative Deaths.

Error rates/quality of perioperative care

Improving perioperative care was a specific goal of SPI2. We selected patients undergoing two types of major surgical operation (total hip replacement and open colectomy) for study. We developed a set of

Box 3: Standards (criteria) for explicit review of perioperative care

- Administration of prophylactic antibiotics before incision
- The use of prophylactic deep vein thrombosis treatment (unless contraindicated), which included pharmacological intervention (unfractionated or low molecular weight heparin) or mechanical interventions (such as antithromboembolism stockings, foot pumps, and sequential compression devices) or both
- Intraoperative monitoring of temperature (on at least one occasion)
- The use of "advanced methods" of pain control (epidural anaesthesia or analgesia controlled by patient, or both)

explicit criteria for perioperative care (box 3) based on IHI "bundles" (collections of carefully packaged evidence based standards directed at a particular condition or clinical scenario) and published clinical guidelines.[4-7] Case notes were obtained from the nine control and nine SPI2 hospitals. We used one epoch before the intervention (epoch 2, October 2006 to March 2007) for comparison with the epoch after the intervention (epoch 3, October 2008 to March 2009). We planned to analyse 10 sets of case notes from each epoch (five of each type of surgical operation) to yield a total sample of 360. To control for seasonal effects, the case notes were spread across each time period (two a month). The notes were processed (including anonymisation) in the same way as the acute medical notes (see above).

A medically trained reviewer (UN) reviewed the case notes. The first 20 cases were read jointly by UN and RL and discussed for training purposes. The notes were partially scrambled over epochs to assess, and if necessary control for, learning/fatigue effects. Agreement between raters was measured by using 27 sets of case notes reviewed by a second reviewer (AK), a surgical trainee. The sample was sufficient to detect a 25 percentage point effect of SPI at 80% power given a baseline rate of 50%.

Indirect measure of hand hygiene

Improving hand hygiene was a specific aim of SPI. A separate UK initiative to improve hospital hand hygiene—the "cleanyourhands" campaign[8]—was also rolled out in England and Wales between December 2004 and June 2005 and continued in subsequent years. It sought to make alcohol hand rub available at the bedside, as well as posters on wards updated monthly, and encouraged patients to ask staff to clean their hands. We tested the hypothesis that SPI2 would have an additional effect over that of "cleanyourhands."

We collected monthly data from NHS Logistics on consumption of soap and alcohol hand rub (as an indirect measure of compliance with hand hygiene) as an extension of the National Observational Study to Evaluate the "Cleanyourhands" campaign (NOSEC).[9] Data were collected on a monthly basis from July 2004 to September 2008. This spanned a period before SPI (July 2004 to February 2007) and a period concurrent with the intervention (March 2007 to September 2008). To adjust for potential variations in consumption because of hospital size, these data, which were available at hospital trust level, were expressed as a rate (in litres) per 1000 bed occupancy days. Bed occupancy days were based on yearly averages spanning financial years.[10]

Outcomes

Adverse events—SPI2 aimed to reduce adverse events by 30%. We identified adverse events using the holistic review of acute medical case notes (see holistic review above), and we assessed the degree of preventability, as in the evaluation of SPI1.[1]

Mortality among acute respiratory patients—SPI2 sought to reduce hospital mortality by 15%. We compared mortality among acute medical patients whose notes were included in the explicit review from before and after the intervention.

Mortality among patients in intensive care units—As a further check on mortality, we assessed mortality in intensive care using data from the case mix programme.[11] Run by the Intensive Care National Audit and Research Centre (ICNARC), this programme is a comparative audit. Data from the intensive care units in all of the study hospitals were available on a monthly basis for six months before SPI2 (October 2006-March 2007) and for six months after the intervention (October 2008-March 2009). Data were available for the numbers of deaths and expected numbers of deaths, which we then used to calculate observed to expected mortality ratios. Data were also available on two mean risk prediction scores: the APACHE (acute physiological and chronic health evaluation) II score[12] and the ICNARC score[13] for patients admitted directly from a ward. We adjusted for these covariates in the analysis.

Infection control—Several components of the SPI related to infection control. We assessed rates of infection with *C difficile* and MRSA from data from the Health Protection Agency on cases of *C difficile* and diarrhoea associated with MRSA bacteraemia in the study sites. Data on *C difficile* were available on a three monthly basis for the period January 2004 to June 2009. MRSA data were available every three months from April 2001 to September 2009. The data therefore spanned a period before the intervention (April 2001 or January 2004 to March 2007), a period concurrent with the intervention (April 2007 to September 2008), and a period after the intervention (October 2008 to June 2009 or September 2009). As required by the Health Protection Agency, the data were expressed as rates per 1000 for *C difficile* and per 100 000 bed occupancy days for MRSA.

Patients' satisfaction—An improvement in patients' satisfaction was not a specific aim of SPI but might be a feasible outcome if an overall improvement in hospital quality occurred. We analysed data from NHS patient surveys with the same methods used in the SPI1 evaluation. Data were collected in October to December 2006 (before the intervention) and October to December 2008 (after the intervention).

Table 3 | Staff survey scores in control and SPI2 hospitals at two periods in evaluation of phase two of Safer Patients Initiative (SPI2)

	Control hospitals					SPI2 hospitals							
	Survey 1		Survey 2*		Absolute % change	Survey 1		Survey 2*		Absolute % change	Range at baseline‡	Effect of SPI2 (99% CI), P value	Point estimate favours SPI2
Survey question†	No of responders	Score (SE)	No of responders	Score (SE)		No of responders	Score (SE)	No of responders	Score SE)				
Well structured appraisals within previous 12 months[14 15]	3477	28 (1)	3429	28 (1)	−1	3783	28 (1)	3734	26 (1)	−2	20-39	−3 (−9 to 3), 0.191	No
Working in well structured teams[16]	3498	36 (1)	3408	37 (1)	1	3781	38 (1)	3747	38 (1)	0	32-42	−4 (−12 to 4), 0.205	No
Witnessed potentially harmful errors or near misses in previous month	3602	37 (1)	3532	33 (1)	−4	3918	41 (1)	3851	40 (1)	−1	32-47	4 (−3 to 10), 0.167	No
Work related injury in previous 12 months	3524	19 (1)	3490	16 (1)	−3	3848	19 (1)	3796	18 (1)	−1	16-23	2 (−2 to 5), 0.182	No
Work related stress in previous 12 months	3575	33 (1)	3532	27 (1)	−6	3882	32 (1)	3842	27 (1)	−6	26-40	1 (−5 to 6), 0.670	No
Physical violence from patients/relatives in previous 12 months	3598	11 (1)	3536	11 (1)	−1	3884	11 (1)	3849	11 (1)	0	7-16	1 (−3 to 3), 0.645	No
Intention to leave[17]	3557	3.26 (0.02)	3544	3.40 (0.02)	0.14	3880	3.31 (0.01)	3865	3.42 (0.01)	0.11	3.07-3.50	−0.04 (−0.12 to 0.04), 0.198	Yes
Staff job satisfaction[17]	3593	3.34 (0.01)	3568	3.44 (0.01)	0.10	3902	3.40 (0.01)	3898	3.49 (0.01)	0.09	3.23-3.50	−0.02 (−0.08 to 0.04), 0.422	No
Quality of work-life balance[17]	3568	2.77 (0.02)	3536	2.56 (0.02)	−0.22	3868	2.68 (0.02)	3857	2.51 (0.02)	−0.17	2.46-2.97	0.05 (−0.04 to 0.14), 0.142	Yes
Support from supervisors[17]	3583	3.39 (0.02)	3551	3.56 (0.02)	0.17	3894	3.43 (0.01)	3869	3.61 (0.01)	0.18	3.22-3.53	0.00 (−0.08 to 0.07), 0.889	—
Organisational climate[17 18]	3578	2.79 (0.01)	3551	2.87 (0.01)	0.08	3861	2.91 (0.01)	3886	2.92 (0.01)	0.01	2.52-3.07	−0.07 (−0.14 to 0.00), 0.009	No
Fairness and effectiveness of incident reporting procedures[17]§	3555	3.36 (0.01)	3487	3.41 (0.01)	0.05	3861	3.41 (0.01)	3803	3.45 (0.01)	0.04	3.27-3.54	−0.01 (−0.05 to 0.04), 0.664	No
Availability of hand washing materials[17]§	2939	4.58 (0.01)	3126	4.75 (0.01)	0.17	3231	4.51 (0.01)	3418	4.67 (0.01)	0.16	4.32-4.72	−0.01 (−0.07 to 0.04), 0.587	No

*After intervention.
†First six scores are percentages, simply reflecting percentage of respondents who answered "yes" to single question or set of questions. The seven others are on scale of 1-5 and are based on mean of between three and six questions, each of which was scored between 1 and 5 for each respondent. For six of these seven scores, higher scores are better, though for "intention to leave" lower scores are better.
‡Indicates range of scores across intervention and control hospitals in first survey to give some context for level of change shown. Difference in change and corresponding confidence interval does not necessarily reflect difference in absolute change because of inclusion of covariates in models tested.
§These scores were not included in SPI1 evaluation.

Table 4 | Medical history taking (% of patients asked required questions) before (epoch 1) and after (epoch 2) phase two of Safer Patients Initiative (SPI2) and effect of SPI. Figures are percentages (binomial standard errors (SE)) and odds ratios (99% confidence intervals) and P values for effect of SPI2

	Control hospitals			SPI2 hospitals			
	Epoch 1 (n=120)	Epoch 2 (n=123)	Epoch 3* (n=112)	Epoch 1 (n=116)	Epoch 2 (n=117)	Epoch 3* (n=114)	OR (99% CI)†, P value
Duration of "presenting" symptom	93 (2)	91 (3)	96 (2)	97 (2)	98 (1)	99 (1)	1.7 (0.07 to 40.3), 0.672
Normal exercise tolerance	27 (4)	32 (4)	38 (5)	39 (5)	38 (5)	34 (5)	0.7 (0.3 to 1.7), 0.312
Presence/absence shortness of breath	88 (3)	91 (3)	88 (3)	91 (3)	93 (2)	92 (3)	1.3 (0.3 to 5.7), 0.701
Presence/absence orthopnoea	23 (4)	28 (4)	17 (4)	33 (4)	29 (4)	18 (4)	0.9 (0.3 to 2.6), 0.749
Presence/absence cough	88 (3)	89 (3)	87 (3)	91 (3)	92 (3)	84 (4)	0.7 (0.2 to 2.4), 0.407
If cough present, was it productive	78 (4)	85 (3)	78 (4)	87 (3)	88 (3)	77 (4)	0.7 (0.2 to 2.1), 0.418
Smoking history taken	74 (4)	81 (4)	66 (5)	78 (4)	80 (4)	74 (4)	1.5 (0.5 to 4.0), 0.313
Presence/absence of haemoptysis	22 (4)	28 (4)	16 (4)	25 (4)	23 (4)	26 (4)	2.2 (0.7 to 6.5), 0.061
Chest pain (of any type)	68 (4)	72 (4)	55 (5)	54 (5)	66 (4)	60 (5)	2.1 (0.9 to 5.2), 0.028
Occupation/previous occupation	44 (5)	38 (4)	54 (5)	35 (5)	39 (5)	38 (5)	0.6 (0.3 to 1.5), 0.178
Pets at home	3 (2)	3 (2)	1 (1)	2 (1)	3 (2)	6 (2)	8.3 (0.3 to 210.0), 0.093
% over all items	56	58	54	68	59	57	—

*After intervention.
†OR >1 favours SPI2.

Statistical methods

When necessary we modified the methods and models used in the SPI1 analysis to accommodate the presence of an additional epoch before the intervention. The temporal change between epochs 1 and 2 was included as a fixed effect in the statistical models. As before, the effect of the SPI was identified as the discrepancy between the two arms of the study in the change experienced from the periods before and after the intervention. All patient level analyses were adjusted for age and sex. When the data were a time series (as opposed to the before and after data available in SPI1), we used population averaged models (fitted with generalised estimating equations), incorporating hospital level random effects. Here cubic polynomials were used to model temporal effects, with a third order autoregressive correlation structure for the residuals. In this context, the effect of SPI2 was assessed as an interaction between time and study arm. An exchangeable (rather than autoregressive) correlation structure was used for the analysis of mortality data from intensive care units, as this was available only for two non-contiguous six month time slots, before and after the intervention. All analysis was carried out in Stata v10.

RESULTS

Staff survey

In the nine SPI2 hospitals, the overall response rate in the "before" survey was 53% (3957 of 7402 valid questionnaires returned). This rate remained the same for the "after" survey (3940/7448). In the nine control hospitals, the response rates were 50% (3634/7301) and 49% (3616/7424), respectively.

For only one of the 13 scores (organisational climate) was there a significant (P<0.01) change over time between the control hospitals and SPI2 hospitals (table 3), which favoured the controls. The effect size for the difference in change between the control and SPI2 hospitals after adjustment for covariates was modest, at 0.07 points on a 5 point scale, where the range between hospitals at baseline was 0.55 points.

Error rates/quality of acute medical care

Explicit review

SPI2 hospitals yielded 347 sets of case notes; 355 were obtained from controls (table 4). We found a significant reviewer learning/fatigue effect (P=0.009) in the review of prescribing errors, with a decreasing rate of error detection with time of review; this was controlled for in the analysis. Despite masking, the reviewer became aware of the patient's name in 1.1% (8) of

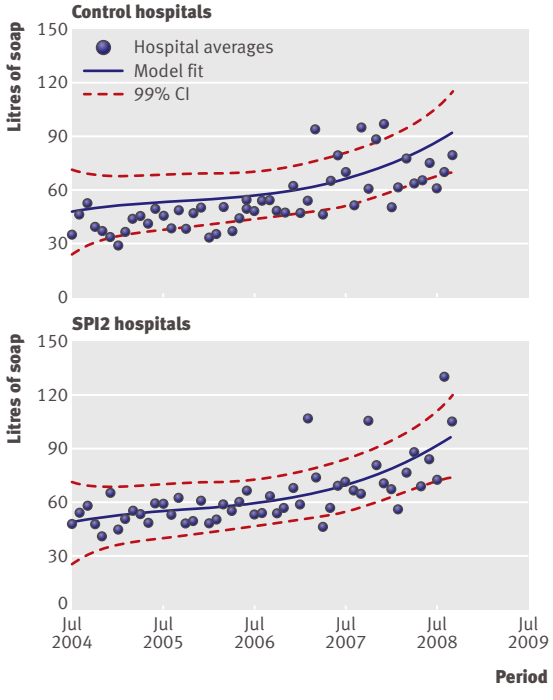

Fig 1 | Rates of consumption of soap over time in control and SPI2 hospitals

Table 5 | Vital signs in phase two of Safer Patients Initiative (SPI2). Figures are percentage compliance with standards with standard errors (SE) and odds ratios for changes over time and effect of SPI2

	Control hospitals			SPI2 hospitals			OR (99% CI)†, P value	
	Epoch 1 (n=120)	Epoch 2 (n=123)	Epoch 3* (n=112)	Epoch 1 (n=116)	Epoch 2 (n=117)	Epoch 3* (n=114)	Changes in controls	Effect of SPI2
On admission								
Temperature	97 (2)	99 (1)	99 (1)	99 (1)	99 (1)	97 (2)	0.7 (0.02 to 24.0), 0.823	0.1 (0.002 to 4.1), 0.108
Respiratory rate	96 (2)	99 (1)	100	97 (2)	98 (1)	100	NA	NA
Cyanosis/oxygen saturation	98 (1)	98 (1)	100	99 (1)	99 (1)	100	NA	NA
Confusion/mental state	53 (5)	72 (4)	74 (4.2)	63 (5)	57 (5)	81 (4)	1.8 (0.8 to 3.7), 0.045	1.7 (0.6 to 4.5), 0.187
Pulse	98 (1)	99 (1)	100	99 (1)	99 (1)	100	NA	NA
Blood pressure	98 (1)	99 (1)	100	99 (1)	99 (1)	100	NA	NA
At 6 hours								
Temperature	62 (5)	70 (4)	70 (4)	63 (5)	78 (4)	68 (4)	1.4 (0.7 to 2.8), 0.239	0.8 (0.3 to 1.9), 0.457
Respiratory rate	41 (5)	69 (4)	72 (4)	47 (5)	76 (4)	78 (4)	2.1 (1.0 to 4.3), 0.010	1.0 (0.4 to 2.8), 0.907
Pulses	69 (4)	73 (4)	75 (4)	65 (5)	81 (4)	80 (4)	1.3 (0.6 to 2.8), 0.327	1.2 (0.4 to 3.3), 0.662
Oxygen saturation	62 (5)	72 (4)	74 (4)	61 (5)	79 (4)	80 (4)	1.4 (0.7 to 3.0), 0.223	1.2 (0.4 to 3.1), 0.703
At 12 hours								
Temperature	58 (5)	71 (4)	69 (4)	59 (5)	70 (4)	73 (4)	1.2 (0.6 to 2.4), 0.583	1.2 (0.5 to 2.9), 0.685
Respiratory rate	35 (4)	70 (4)	73 (4)	45 (5)	68 (4)	79 (4)	2.4 (1.1 to 5.0), 0.002	1.2 (0.4 to 3.1), 0.713
Pulse	63 (4)	76 (4)	75 (4)	60 (5)	71 (4)	80 (4)	1.2 (0.6 to 2.5), 0.510	1.5 (0.6 to 4.1), 0.268
Oxygen saturation	54 (5)	76 (4)	74 (4)	57 (5)	71 (4)	80 (4)	1.4 (0.7 to 2.9), 0.231	1.4 (0.5 to 3.6), 0.430
Routine investigations								
Urea and electrolytes	99 (1)	98 (1)	99 (1)	100	99 (1)	100	0.6 (0.01 to 27.7), 0.762	NA
Chest x ray	97 (2)	98 (1)	97 (2)	97 (2)	98 (1)	100	0.7 (0.1 to 5.6), 0.641	NA
Full blood count	98 (1)	98 (1)	99 (1)	99 (1)	99 (1)	100	1.7 (0.1 to 40.4), 0.663	NA

NA=not applicable because of 100% in cells.
*After intervention.
†OR >1 favours SPI2. No items showed significant variation between hospitals within arms.

cases; of hospital of origin in 0.4% (3) of cases; and of epoch in 16.7% (117) of cases. Baseline comparisons showed no significant differences between control and SPI2 hospitals.

There was no apparent net additional effect of SPI2 and no significant effect for any of the end points measured (tables 5-7). For two items (exercise tolerance and occupation) measured in relation to history taking, there was significant evidence of an improvement in hospitals over time. This occurred in both control and SPI2 hospitals.[2]

Compliance in taking observations of patients at 6 and 12 hours after admission also improved in both groups of hospitals. This was most evident for respiratory rate, where practice continued to improve across all three epochs. There was a considerable and significant increase over time in use of the CURB score (a clinical prediction rule for predicting mortality from community acquired pneumonia, see table 6) (odds ratio 7.3, 99% confidence interval 1.4 to 37.7), but again differences were not significant between control and SPI2 hospitals. Point estimates for six of the eight standards for monitoring vital signs in the first 12 hours after admission and for use of the CURB score favoured SPI2 hospitals. There were no significant effects of SPI2 either over time or in favour of SPI in quality of prescribing (error rate ratio (estimated from population averaged negative binomial model) 0.9, 0.5 to 1.5; P=0.444) (table 7). Further details, along with tests for homogeneity of baseline end points among control and SPI2 hospitals and the effect of covariates are given in the full report.[2]

Holistic review

A total of 725 sets of case notes were reviewed. In epoch 1 we reviewed 126 sets from control hospitals and 117 from SPI2 hospitals. The corresponding figures were 126 and 120 for epoch 2 and 114 and 122 for epoch 3. Agreement between the principal reviewer and the reliability reviewer was low (κ=0.08).

A single patient could have more than one error. In the control hospitals, the average number of errors per 100 patients decreased over the three epochs from 53 (epoch 1), to 40 (epoch 2), to 31 (epoch 3) per 100 patients. In the SPI2 hospitals, the average number of errors per 100 patients was relatively stable over epochs at 36, 45, and 39 per 100 patients. Differences in changes in the average number of errors before and after the intervention, across control and SPI2 hospitals, were not significant (rate ratio 1.47, 0.74 to 2.90). As in SPI1, diagnostic errors were the most common type of error. In SPI2, however, errors in clinical reasoning were the second most common category, exceeding medication error and hospital acquired infections (see full report).[2]

Error rates/quality of perioperative care

We retrieved 242 sets of notes; 127 were from admissions for total hip replacements and 115 from admissions for open colectomies. One person reviewed

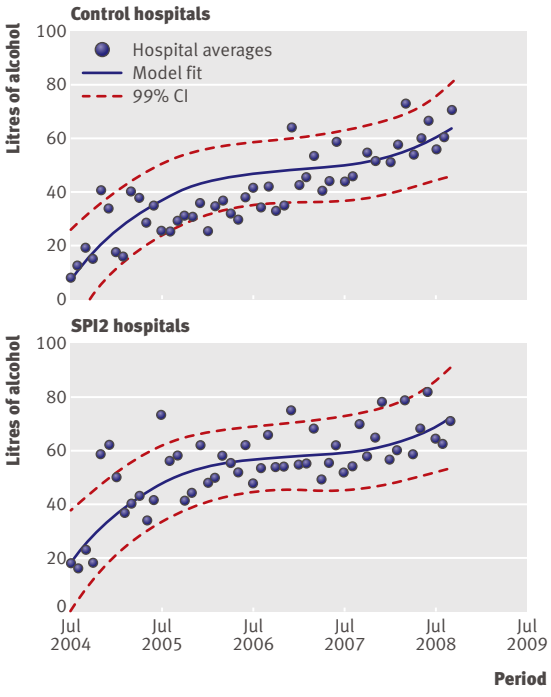

Fig 2 | Rates of consumption of alcohol hand rub over time in control and SPI2 hospitals

these; a second reviewer also examined a sample of 27. Percentage agreement was 93% for antibiotics and 96% for deep vein thrombosis prophylaxis, but κ values could not be calculated because one reviewer put all cases in the same category. Agreement was 85% and 59% for pain relief and temperature monitoring, with corresponding κ values of 0.46 and 0.24. The hospitals were similar at baseline, except with respect to intraoperative monitoring of temperature, where controls had more room for improvement.

There was little room for improvement for two of the four standards examined. For antibiotic prophylaxis, compliance exceeded 94% in all cases; for deep vein thrombosis prophylaxis, compliance exceeded 98% in all cases. Intraoperative monitoring of temperature improved over time in both groups, although this result was not significant (odds ratio 1.8, 0.4 to 7.6; P=0.279). There were no significant SPI2 effects for any of the four clinical standards examined (table 8).

Indirect measure of hand hygiene

Data on consumption of soap and alcohol hand rub were available from epochs 1 and 2 for nine and eight of the control trusts and for seven and six of the SPI2 trusts, respectively. Rates of consumption increased in both control and SPI2 hospitals over the study period (figs 1 and 2), suggesting considerable improvement. The rate of increase in consumption (that is, differences in differences), however, was not significantly greater in the SPI2 hospitals than in the control hospitals (P=0.760, favouring controls, and P=0.889, favouring SPI2, respectively).

Outcomes

Adverse events in patients on acute medical wards

The holistic review identified adverse events among patients on acute wards. Reliability between raters was no better than chance (κ=0.0). Over all hospitals and all epochs, the principal reviewer identified 22 adverse events among the 725 sets of case notes, giving an average adverse event rate of 3.03 per 100 patients. In the control hospitals, the average number of adverse events per 100 patients decreased over the three epochs, from 4.8 per 100 patients in epoch 1, to 4.0 (epoch 2), and 3.5 (epoch 3). In contrast, in the SPI2 hospitals the average number of adverse events per 100 patients increased from 0.9 per 10 in epoch 1 to 5.0 per 100 in epoch 2, but decreased to zero in epoch 3. Again, differences in the change in numbers of adverse events across control and SPI2 hospitals were not significant (rate ratio 1.47, 0.74 to 2.90).

The principal reviewer identified strong or certain evidence of preventability in four of the 22 adverse events (that is, 0.5% of cases overall). These four adverse events occurred in the epochs before the intervention (see the companion paper[1]).

Mortality in acute medical patients in case note review

Ninety seven of the 702 patients included in the explicit review of acute medical care died (14%) (table 9). At baseline crude mortality was higher in the control hospitals than in the SPI2 hospitals (odds ratio 0.7, 0.2 to 2.1; P=0.391). Neither this, nor any other effect, including that of the SPI, was significant at the predetermined 1% level after adjustment for age, sex, and number of comorbidities (0.3, 0.08 to 1.4), although the result was just significant (P=0.043) at the 5% level. The odds ratio

Table 6 | Use of systemic steroids, CURB score, and other standards applicable to specific cases—compliance with standards in phase two of Safer Patients Initiative (SPI2). Figures are numbers (percentage, SE) and odds ratios (99% confidence interval) and P values for effect of SPI2

	Control hospitals			SPI2 hospitals			
Standard	Epoch 1	Epoch 2	Epoch 3*	Epoch 1	Epoch 2	Epoch 3*	Effect of SPI2†
Asthma or COPD: steroids given within 24 hours	70 (84, 4)	63 (92, 4)	56 (93, 4)	59 (92, 4)	74 (93, 3)	53 (94, 3)	0.6 (0.05 to 6.8), 0.568
Asthma: peak flow recorded	10 (80, 13)	11 (64, 15)	5 (40, 22)	24 (79, 8)	18 (94, 5)	8 7 (5, 15)	29.7 (0.1 to 16000), 0.165
Community acquired pneumonia: CURB score recorded	52 (2, 2)	67 (22, 5)	56 (21, 6)	50 (2, 2)	44 (25, 6)	60 (42, 6)	2.1 (0.4 to 11.1), 0.236

COPD=chronic obstructive pulmonary disease; CURB=confusion/urea/respiratory rate/blood pressure score.
*After intervention.
†OR >1 favours SPI2. No items showed significant variation between hospitals within arms.

Table 7 | Analysis of prescribing errors before (epoch 1) and after (epoch 3) phase two of Safer Patients Initiative (SPI2)*

	Control hospitals			SPI2 hospitals		
	Epoch 1	Epoch 2	Epoch 3†	Epoch 1	Epoch 2	Epoch 3†
No of patients‡	120	122	112	113	117	114
No of prescriptions	2953	3269	2871	2529	2938	2656
Prescriptions per patient	24.6	26.8	25.6	22.4	25.1	23.3
No of errors	345	298	216	251	266	167
Error rate (SE) per prescription	0.12 (0.02)	0.09 (0.01)	0.08 (0.01)	0.101(0.02)	0.09 (0.01)	0.06 (0.01)

*Breakdown of error types, including failure to reconcile patient's previous medicines with prescription on admission (particular focus of SPI), available in full report (www.haps.bham.ac.uk/publichealth/psrp/EvalSPI.shtml).2
†After intervention.
‡With medication charts available for review.

for change in controls was 1.4 (0.06 to 3.1; P=0.320). Sex and number of comorbidities were included as patient level covariates, but only age was significant (P<0.001). The mortality rate increased by 10.3% (6.8% to 15.1%) per year of patient age.

In the holistic review we reviewed notes from 725 patients (compared with 702 in the explicit review). Of these, 91 (13%) died. The principal reviewer (MC) found potentially preventable factors in six of these cases and the second reviewer (MCD) also found six; five cases were common to both reviewers. The second reviewer was more "hawkish," however, placing five of her six cases in the category of 50% or greater likelihood of preventability—that is, she concluded that it was more likely than not that death during the hospital stay could have been prevented had the putative failure in care not occurred. The principal reviewer was "dove-like," placing all of his cases in the "less than 50%" category. Table 10 gives a breakdown of deaths by level of preventability and reviewer. The total percentage of high (>50%) preventability deaths common to both reviewers was 5% (5.5% of deaths). They identified seven (7.7% of deaths) deaths with any potential preventable factors.

Mortality in intensive care units

Seven control and seven SPI2 hospitals supplied data to ICNARC for the period before the intervention period and six control and eight SPI2 hospitals for after the intervention. Based on length of stay, intensive care units in control hospitals might have been dealing with a different case mix to the SPI2 hospitals. APACHE and ICNARC scores were similar, however, and not significantly different between groups and over time. The rate of observed to expected mortality increased in the control hospitals and decreased in the SPI2 hospitals over the study period (table 11). The adjusted difference in difference was not significant (P=0.250).[2]

Infection control: rates of C difficile and MRSA

Data on the numbers of cases of *C difficile* and MRSA were available for all 18 trusts. The infection rate for *C difficile* decreased over the study period in both the control and SPI2 hospitals. The point estimate favours SPI2 hospitals, but differences in changes were not significant between control and SPI2 hospitals (P=0.652). Figure 3 shows the smoothed estimated rates of *C difficile* infection per 1000 bed occupied days by control and SPI2 hospitals.

The median infection rate of MRSA also decreased over the study period in both the control and SPI2 hospitals. Again, it was not possible to detect an effect of SPI in this improvement: differences in changes were not significant between control and SPI2 hospitals (P=0.693) although the point estimate favours SPI2. Figure 4 outlines the estimated smoothed rates of MSRA infection per 100 000 bed occupied days by control and SPI2 hospitals.

Survey of patients

For the first survey, the overall response rate was 62% (4328 of 7010 valid questionnaires returned) in the nine SPI2 hospitals; for the second it was slightly lower at 55% (3762/6810). In the nine control hospitals, the response rates were 63% (4262/6791) and 57% (3973/6913), respectively. Table 12 shows the changes

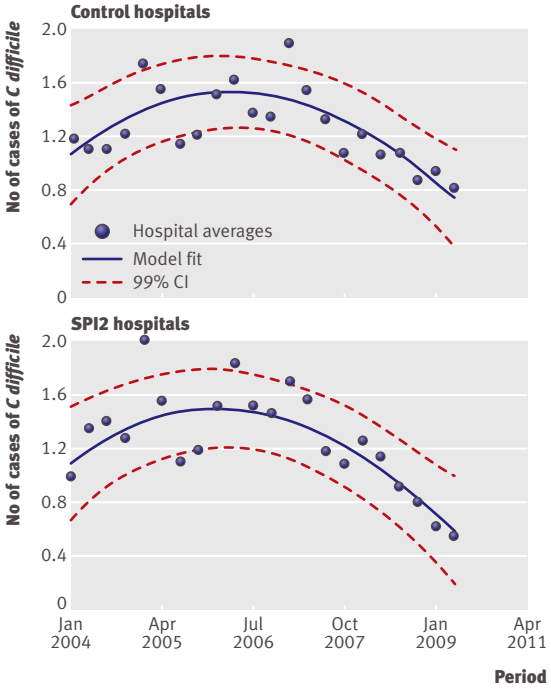

Fig 3 | Rates of cases of *C difficile* per 1000 bed days in control and SPI2 hospitals

Table 8 | Rates of compliance with perioperative care standards before (epoch 2) and after (epoch 3) phase two of Safer Patients Initiative (SPI2). Figures are percentages (standard error (SE)) and odds ratios and P values for effect of SPI2

	Control hospitals		SPI2 hospitals		
	Before intervention (n=51)	After intervention (n=43)	Before intervention (n=79)	After intervention (n=69)	OR (99% CI)*, P value
"Advanced method" of pain relief†	94 (4)	95 (4)	85 (4)	83 (5)	0.8 (0.03 to 18.4), 0.820
Perioperative antibiotic given	94 (3)	100	98 (2)	97 (2)	NA
Temperature monitored‡	16 (5)	30 (7)	29 (5)	41 (6)	0.9 (0.1 to 5.2), 0.854
Appropriate DVT prophylaxis§	100	100	99 (1)	100	NA

DVT=deep vein thrombosis; NA=not applicable because of 100% in cells.
*OR >1 favours SPI2.
†Hospital staff identified 15 cases with contraindications to this standard, all of which were corroborated by reviewers. Data relate to 227 patients who were eligible.
‡ Evidence of heterogeneity between hospitals at baseline.
§ Three patients had contraindications yielding denominator of 238. Withheld in only two patients with no contraindications but wrongly administered in two patients with contraindication.

in both control and SPI2 hospitals on each of the five scores identified, along with the differences between the groups in these changes and associated 99% confidence intervals. All five scores improved over the study period in both the control and SPI2 hospitals; none showed any significantly different changes between the two groups.

DISCUSSION

Commentaries on patient safety in the United States five years after the publication of two key reports on patient safety in 2000 were characterised by some despair at an apparent lack of progress.[19] Our data suggest that a more encouraging story on patient safety in the NHS can now be told.

Baseline performance across hospitals was already high on many criteria relating to quality, leaving little room for improvement. Over 90% of patients with an acute exacerbation of obstructive airways disease received steroids when indicated, and rates of perioperative prophylaxis against venous thrombosis and infection approached 100%, corroborating an earlier study.[20] Where scope for improvement existed, we found many examples of improved, and none of worsening, practice. Vigilance in relation to monitoring vital signs on acute medical wards and use of severity scoring, observed in our study of SPI1, continued to improve. A strong upward trend in recording intraoperative temperature was noted. Rates of handwashing seemed to have increased significantly, and the incidence of *C difficile* and MRSA infection fell. Though results of the staff survey showed little change over time, the survey of patients showed improvement across all five prespecified dimensions, suggesting a better experience for patients. There was even an improvement in medical history taking. Adverse event rates (3.03% in our study) seemed similar to those reported in the Harvard medical practice study (3.7%), which was based on data collected in 1984.[21] We found low levels of preventability among adverse events overall (about 20%) and among deaths (less than 10%). If these findings are corroborated, they have implications for future evaluations and performance management, as the signal (preventable adverse events) seems to be buried in a lot of noise (non-preventable adverse events).

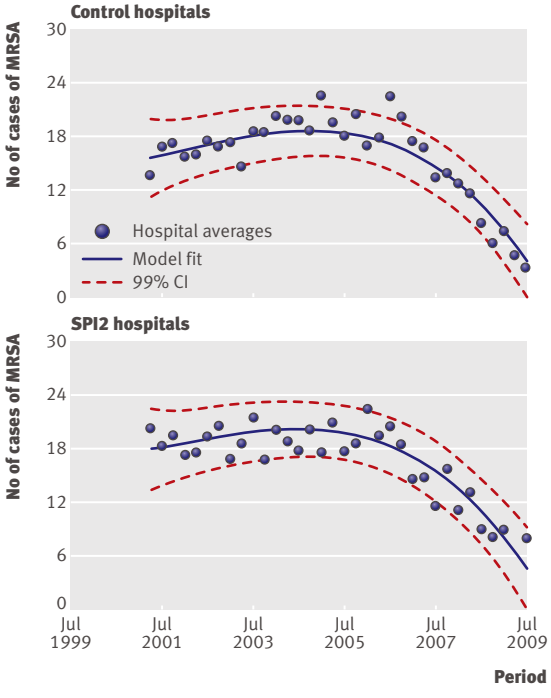

Fig 4 | Rates of cases of MRSA per 100 000 bed days in control and SPI2 hospitals

The data we collected on SPI2 suggest that an additional effect of SPI is difficult to detect over and above the improvements occurring across the health service generally. Indeed, in a reversal of our evaluation of SPI1, organisational climate as measured by the staff survey favoured the controls. Adherence rates for many of the specific criteria reflecting quality of care remained high over time in both groups of hospitals, possibly reflecting a long history of quality improvement in areas such as perioperative care. Those areas that underwent marked improvement did so to a similar degree in both sets of hospitals. One exception was the drop in mortality among acute medical cases in SPI2 hospitals and an unexplained rise in control hospitals, such that the difference in differences would

Table 9 | Mortality among acute medical care patients whose case notes were explicitly reviewed before (epoch 1) and after (epoch 3) phase 2 of Safer Patients Initiative (SPI2)

	Control hospitals			SPI2 hospitals		
	Epoch 1	Epoch 2	Epoch 3*	Epoch 1	Epoch 2	Epoch 3*
No of patients	120	123	112	116	117	114
Deaths	18	24	24	9	15	7
% mortality (SE)	15 (3.3)	20 (3.6)	21 (3.9)	8 (2.5)	13 (3.1)	6 (2.3)
Mean (SD) age (years)	77.6 (7.7)	81.1 (7.9)	79.6 (8.0)	77.7 (7.6)	78.1 (7.1)	80.6 (7.8)
% women	63.3	53.7	53.6	53.4	50.4	52.6
Mean No of comorbidities	2.9	3.1	2.6	2.8	3.0	2.9

*After intervention.

have been just significant if we had selected a $P<0.05$ threshold. Any suggestion of a difference in mortality rates resulting from difference in the quality of care, however, does not align well with the review of the quality of care observed among those same case notes. The observed difference in overall mortality cannot be accounted for by a difference in preventable deaths as only seven of the 91 deaths fell into the possibly preventable category. Overall, though there is considerable evidence of good or improved quality and safety in NHS hospitals, we could not detect a net effect attributable to SPI2 with our study measures. This largely mirrors the evaluation of SPI1,[1] though the latter did show an effect of SPI on the quality of monitoring the respiratory rate. Table 13 summarises the effects of both phases of SPI versus control, in terms of direction of the point estimate and degree of significance.

Strengths and weaknesses of this study

The argument we made in the companion article,[1] and elsewhere,[3] that studies of quality improvement interventions should follow predefined protocols and incorporate contemporaneous controls is reinforced by our study of SPI2, where many end points improved significantly across both SPI2 hospitals and controls. We have also shown the importance of using a difference in difference approach to analysis to overcome the ambiguities of single difference studies. This method is widely used in economics research, where there are policy and other changes occurring during the implementation of a programme,[22] and it is clearly suitable for evaluations of quality improvement programmes in healthcare. We have also shown the need to allow for learning/fatigue effects in reviewing.

A particular strength of our study arises from its possibilities for triangulation. While available funding did not permit us to build further qualitative studies into the design, we did have various internal controls. Findings on the use of handwashing materials and rates of two different types of infection support the hypothesis of general improvement in this area. The observation that vital signs were recorded with increasing diligence and that risk scoring was used more often supports the idea that patients at risk of deterioration were being monitored more diligently. Mortality rates on acute medical wards could be triangulated, not only by an audit of compliance with process standards, but also by scrutinising each death in the sample to see if it could have been caused by poor care.

With hindsight, there are some things that we would do differently in this study. We would not measure all prescribing errors as this is expensive, and many errors are minor and of uncertain validity as surrogates for

Table 10 | Preventable deaths among acute medical care cases where notes were reviewed holistically across study epochs before (epoch 1) and after (epoch 3) phase two of Safer Patients Initiative (SPI2)

Epoch	No of deaths (cases reviewed)	Preventable deaths ≥50%*			Preventable deaths <50%†		
		1st reviewer only	2nd reviewer only	Both reviewers	1st reviewer only	2nd reviewer only	Both reviewers
Control hospital							
1	17 (126)	0	0	1	0	0	1
2	24 (126)	0	0	1	1	0	1
3‡	23 (114)	0	0	2	0	1	2
Total	64 (366)	0	0	4	1	1	4
SPI2 hospitals							
1	9 (117)	0	0	0	0	0	0
2	11 (120)	0	0	1	0	0	1
3‡	7 (122)	0	0	0	0	0	0
Total	27 (359)	0	0	1	0	0	1

*Preventable deaths ≥50%: on balance of probabilities substandard practice led to death.
†Preventable deaths <50%: substandard practice could have led to death but probability that it did was <50%.
‡After intervention.

serious error.[23] We would instead concentrate on errors whose serious nature had been established. The reliability reviewer was new to case note reviews, and although several training sessions took place we would approach this training more systematically in the future. Nevertheless, previous work has also found that reliability for holistic reviews is lower than reliability for explicit reviews.[24]

We are developing and evaluating a novel tool based on review of case notes of patients who die in hospital, where each death is scored on a sliding scale of preventability. The aim is to produce a reliable measure of the proportion of hospital deaths that are preventable.

Possible improvements in clinical areas not studied

Though we had an explicit rationale for the clinical areas in which we focused our study, improvements could have occurred in areas that we did not study, such as ventilator acquired pneumonia and central line infections in intensive care units. If improvements did occur in these areas, it is possible that a greater "dose" of SPI was administered in these settings (for example, more activity by SPI "change agents") or that such settings were more responsive to change than those we studied.

Improvements below the level of statistical detection

The absence of an additive SPI effect detected by our study does not exclude smaller effects that might none the less be cost effective. The threshold in England under which an intervention is judged cost effective is about £30 000 (€35 000; $48 000) per quality adjusted life year (QALY). The SPI would, therefore, need to save fewer than seven lives with a mean duration of five healthy years (ignoring discounting) to justify the SPI1 investment of £775 000 per hospital (and an even smaller magnitude of effect would be cost effective at the smaller costs in SPI2 hospitals). An effect of this magnitude cannot be excluded in a study of any feasible size; with many hundreds of deaths taking place in each hospital in each year the signal would be lost in the noise.[25] None the less, large effects postulated in advance of the study have been excluded, at least in the areas examined. The study was, after all, large enough to detect temporal improvements. The 50% and 30% reductions in adverse events that were aims of SPI1 and SPI2, respectively, were unnecessarily large in the sense that much smaller effect sizes would justify the costs of the intervention.

SPI hospitals might have been less sensitive to the intervention

The study was not randomised, and we cannot exclude the possibility that SPI hospitals as a whole were less sensitive to the intervention than controls. There were few differences at baseline, however, and where there was room for improvement among controls, similar room was available for SPI hospitals. It is also possible that SPI works better in some types of hospital than others.[26] We did not have statistical power to test for such interactions.

Possible suboptimal specification or implementation of SPI

Some of the reasons for the absence of an additional detectable SPI effect might lie in the design and implementation of the programme. While interviews conducted with senior staff in the study of SPI1[1] emphasised the "bottom-up" nature of the intervention, this was not necessarily how it was perceived by most ward staff. Despite the enthusiasm and broad understanding of the principles underlying the SPI at a strategic level, the programme and organisational theories of change might not have been sufficiently explicit. For example, no formal protocol for the intervention was published. There is evidence from the qualitative work in SPI1 that the scale of the SPI task was perceived as huge and demanding of resource. There were also suggestions that there was a need for the programme to be purposefully and actively led in each clinical setting, rather than assuming spontaneous "spread" from one setting to another. More work before the intervention might have identified with more precision how and under what conditions the programme would work best and would have more completely specified the underlying theories.

Optimising design and execution of quality improvement programmes is clearly necessary for many reasons, not least to avert the risk of damaging the credibility of such programmes as a whole. A combination of a more explicit programme theory and organisational theory of change, including better specification of the method of vertical and horizontal spread, might, for example, have explicitly confronted the six "universal challenges" for quality improvement (structural, political, cultural, educational, emotional,

Table 11 | Intensive care outcomes in phase two of Safer Patients Initiative (SPI2). Median and interquartile ranges for control and intervention hospitals, before and after intervention period*

Intensive and critical care outcomes*	Control hospitals		SPI2 hospitals		Difference in difference	
	Before	After	Before	After	Change (99% CI)†	P value
Adjusted mortality ratio	1.14 (0.99-1.32)	1.24 (1.02-1.33)	1.04 (0.90-1.15)	0.97 (0.90-1.15)	0.09 (−0.11 to 0.29)	0.25
Mean LOS (hours)	144 (117-174)	147 (126-185)	102 (82-130)	103 (81-137)	5.86 (−22.78 to 34.50)	0.60
Mean APACHE II score	20.4 (17.7-22.6)	19.0 (17.1-20.8)	21.1 (19.1-23.0)	20.3 (17.8-21.8)	−0.83 (−3.63 to 1.98)	0.459
Mean ICNARC score	22.3 (19.5-26.3)	20.7 (18.0-23.5)	22.6 (21.2-25.3)	22.2 (19.7-25.1)	−2.26 (−6.39 to 1.87)	0.16

LOS=length of stay; APACHE=acute physiological and chronic health evaluation; ICNARC=Intensive Care National Audit and Research Centre.
*Before period is October 2006 to March 2007 and after period is October 2008 to March 2009.
†Change <1 favours SPI2.

Table 12 | Patient survey scores* in control and SPI2 hospitals before (survey 1) and after (survey 2) phase two of Safer Patients Initiative (SPI2)

	Control hospitals					SPI2 hospitals						
	Survey 1		Survey 2		Absolute % change	Survey 1		Survey 2		Absolute % change	Range at baseline	Effect of SPI2 (99% CI)†‡, P value
	No of patients	Score (SE)	No of patients	Score (SE)		No of patients	Score (SE)	No of patients	Score (SE)			
Overall, how would you rate the care you received?	4200	82 (0.4)	3913	85 (0.3)	4	4277	80 (0.4)	3705	84 (0.3)	4	75-87	1 (−1 to 3), 0.292
Overall, did you feel you were treated with respect and dignity while you were in the hospital?	4111	78 (0.4)	3807	82 (0.4)	4	4167	76 (0.4)	3604	80 (0.4)	3	65-85	0 (−2 to 2), 0.702
How would you rate how well the doctors and nurses worked together?	4182	87 (0.4)	3878	88 (0.4)	1	4220	88 (0.4)	3677	89 (0.4)	1	83-91	0 (−2 to 2), 0.597
In your opinion, how clean was the hospital room or ward that you were in?	4113	75 (0.4)	3870	77 (0.4)	2	4201	77 (0.4)	3645	78 (0.4)	1	70-80	−1 (− to 1), 0.141
How clean were the toilets and bathrooms that you used in hospital?	4141	76 (0.4)	3877	78 (0.4)	2	4220	78 (0.4)	3665	79 (0.4)	1	70-82	−1 (−3 to 1), 0.204

*Mean values of five survey scores in control and SPI2 hospitals for first and second patient surveys, rated between 0-100.
†Values >1 favours SPI2.
‡Values differ from simple subtraction because of rounding.

and physical/technological),[27] and it might have focused more attention on ensuring clinical engagement and use of clinical networks. Such an approach might have encouraged an earlier recognition that the intervention was broad relative to resources and might have identified that effects were likely to be localised in response to "dose" of intervention. In that case a more focused and less ambitious intervention, and somewhat narrower evaluation, might have been a better strategy.

There is also an argument that participation in SPI could secure greater long term commitment to quality and safety in participating hospitals and that improvements made in the intervention hospitals will either surface at a later date or be sustained better. This hypothesis can be tested only with further data collection, but it is possible that any effect of SPI might be in the form of "stickiness"; intervention hospitals might potentially be better equipped to show sustained improvements after the policy spotlight has moved elsewhere.

Contemporaneous policy and professional forces in the control environment

SPI coincided with a period of unprecedented increase in NHS funding that could have contributed to many of the improvements observed. An important reason for the absence of an additional effect of SPI might lie in the extent of the policy level programmes and initiatives that were largely contemporaneous with the SPI, shared some of its goals, principles, and methods, and acted forcefully on the control environment. For example, the "cleanyourhands" campaign promoted the same goal of improved hand hygiene as the SPI and began around the same time. In addition, the Health Act 2006 introduced new legislation on mandatory requirements for prevention and control of infections associated with healthcare and is likely to have exerted further pressures on hospitals.

Perhaps most importantly, several initiatives with features similar to IHI-style techniques and principles had increasing impact on policy at around the time that the SPI (which was mentored by the IHI) was launched. For example, the Department of Health's Saving Lives programme, beginning in June 2005 with a revised version in 2007,[28] included a self assessment tool for trusts to assess their managerial and clinical performance and a set of "high impact interventions" that were similar to the IHI bundles and were aimed at several clinical processes also targeted by the SPI. The interest in IHI-like interventions might indeed have been prompted or inspired at least in part by the SPI; the House of Commons committee report on patient safety, for example, lists the SPI's beginning in 2004 as among the important policy developments in the patient safety timeline.[29] It is also relevant that many of these policy initiatives had already been anticipated by consensus in professional societies and medical colleges, and thus enjoyed considerable professional legitimacy—a crucial factor in promoting safe and effective practice.[30] Patient safety and quality improvement was also, during the period when SPI was being implemented, drawing increasing attention from journals, professional meetings, and conferences. The SPI programme was thus being implemented at a time when the momentum towards quality improvement was accelerating and when it might itself have been one of the forces implicated in the momentum.

Given that many of the changes in practice being urged at a policy level were so similar to the SPI, and the resource directed at the SPI was relatively small, the SPI itself might not have been a sufficient additional "dose" to generate further detectable differences in participating hospitals: from £270 000 to £775 000 spent over 18 months in hospitals with annual budgets of £150m to £300m might simply be too small. This is perhaps most vividly illustrated by the disappearance

Table 13 | Summary of directions of effects of SPI across all quantitative evaluations of SPI1 and SPI2. Significant results are indicated

	SPI1 *v* control			SPI2 *v* control		
	Outcomes favouring SPI	Outcomes favouring control	Outcomes with no difference	Outcomes favouring SPI	Outcomes favouring control	Outcomes with no difference
Staff survey	7*	3	1	2	10*	1
Vital signs	9*	3	2	7	2	4
Routine investigation	0	3	—	—	—	3
Specific standards	0	3	—	2	—	—
Prescribing	—	1	—	1	—	—
Medical history	5	5	1	6	5	1
Mortality among case notes	1	—	—	1	—	—
Holistic errors	1	—	—	—	1	—
Holistic adverse events	1	—	—	—	1	—
Patient survey	4*	1	—	1	2	2
Intensive care mortality	NA	NA	NA	1	—	—
Handwashing materials:						
Soap	NA	NA	NA	—	1	—
Alcohol hand rub	NA	NA	NA	1	—	—
C difficile rates	NA	NA	NA	1	—	—
MRSA rates	NA	NA	NA	1	—	—
Totals	28	19	4	24	22	11

NA=Not applicable

*One of these results was significant at P<0.01. Across whole study, four of 108 measured outcomes yielded significant results at P<0.01 when changes in SPI were compared with changes in control. Three of these favoured SPI and one favoured controls. Four further results were significant at P<0.05, but not P<0.01; and all favoured SPI (three in SPI1 and one in SPI2).

in SPI2 of the positive impact on measures of compliance with monitoring and response to vital signs deterioration that we found in SPI1. This probably occurred because guidelines on recognition and response to acutely ill patients were issued by the National Institute for Health and Clinical Excellence (NICE) in 2007,[31] just as SPI2 was getting started. The detectable effects of SPI could have been muted compared with a situation where no similar policy changes were occurring.

In clinical research it has long been known that outcomes tend to improve over time, with the result that before and after studies systematically exaggerate treatment effects compared with studies with contemporaneous controls.[32] In clinical research temporal trends are usually the result of various factors apart from the intervention of interest, although exceptions exist—for example, HIV drugs and prostate specific antigen screening diffused into widespread use before evaluations were complete.[33 34] This risk of pre-evaluation diffusion is arguably greater in the case of management interventions that are multi-faceted, not easily containable, and are promoted as part of "continuous improvement" strategies. While new medicines are generally evaluated before they can be licensed and adopted, service interventions can more easily come into general use and generate social reinforcement before a formal evaluation has been put in place. Indeed growing interest in the intervention might be the stimulus both for increasing adoption and for the evaluation. The evidence provided above suggests that something like that happened with SPI and might have occurred in the provocatively null result of the MERIT study of rapid response service on medical wards.[35]

Our results suggest the occurrence of a phenomenon where the measured effect of an intervention is attenuated by similar changes happening more generally. This should be distinguished from the phenomenon of contamination, where the control group receives (some of) the intervention targeted at the study group.[36] In the case of the SPI, "contamination" is an inappropriate descriptor as the study was "anamnestic" and controls were selected after the SPI had been put in place. SPI implementation was well under way when controls were selected and the controls were not exposed to the extensive and expensive mentoring process that SPI entailed. We propose, rather, that a "rising tide" phenomenon was at work; both control and SPI sites were subject to the same tidal forces, and these same latent factors were the source of both a change in practice and the perceived need to evaluate these changes. Under these circumstances it is still worth evaluating an intervention, but this is more akin to evaluating "dose" in clinical research; the idea is pragmatic and aims to find out whether the marginal gains of an extra "push" is worth marginal expenditure or, at least, to provide some evidence to inform such a judgment.

Conclusions

Our studies show encouraging signs of improvements in quality and safety in the NHS in England, but detected only one specific improvement as a result of SPI, and that was confined to the first phase of the

WHAT IS ALREADY KNOWN ON THIS TOPIC

There are many examples of evaluations of interventions to improve the quality of specific clinical processes, but fewer attempting evaluations of system-wide change in whole hospitals

The second phase of an attempt to effect system-wide change, the Safer Patients Initiative, was rolled out in 10 hospitals in England and 10 hospitals in other countries of the UK from March 2007 to September 2008

WHAT THIS STUDY ADDS

Patient safety has improved across the NHS on many of the measures used in our study of English hospitals

No additional effect of the Safer Patients Initiative could be detected

Several possible explanations for the absence of an additional effect of the programme can be offered, including a "rising tide" phenomenon where improvements in patient safety were driven by common forces across the NHS

programme. Any detectable effects of such interventions might take time to surface. Such interventions are likely to benefit from clarity about the theories of change underlying the programme, recognition of the scale of resource and organisational support required to make patient safety efforts work, and improved understanding of how practitioners, middle managers, and organisational systems can be better supported in the face of daunting complexity and multiple priorities. Robust methods are needed to make appropriate conclusions about the impact of quality improvement efforts.

We thank Michael D L Morgan, Martyn R Partridge, and Philip W Ind for their expertise and contribution in the development of the forms for the explicit case note review for respiratory care; Dion Morton and David Thomas for their expertise in the development of the forms for the surgical explicit case note review; Sheldon Stone and Chris Fuller for access to data collected as part of the NOSEC study; David Harrison and Kathy Rowan from ICNARC for access to data collected as part of the Case Mix Programme; Dale Webb, Louise Thomas, and Simona Arena for their help in describing the SPI intervention; Steven Thornton, chief executive of the Health Foundation, for providing a superb role model in sponsoring formal, summative evaluations of service level interventions; Peter Chilton for his assistance in the preparation of this manuscript; and Frank Davidoff, Laura Morlock, and Tim Hofer for excellent comments on the manuscript.

Contributors: AB, MD-W, JD, NB, and RL designed the study and submitted the grant proposal. RL was chief investigator. AB, NB, RL, MG, and BDF designed the forms for the explicit case note review and methods for the explicit case note review. AB, RL, and UN designed the semistructured forms for the holistic case note review and methods for data extraction. AB and UN were responsible for collecting the case note reviews. MG and BDF conducted the review of acute medicine case notes. MC and TN conducted the holistic review of case notes. MC and CD carried out a separate review of deaths. UN and MG designed the database for acute medicine case note review. GR and AB created the queries for data extraction. UN, AB, and RJL designed the forms for the perioperative case note review. AB and UN designed the database for the perioperative case note review. UN and AK conducted the review of case notes. AG analysed all the data from explicit reviews of case note. GR and AB designed and wrote database queries for final analysis and to assess the learning effect on the case reviewers. GR captured processed raw mortality data and calculated hospital standardised mortality rates for hospitals in both arms and undertook analysis of the socioeconomic composition of the admitted patient populations of hospitals in the study. KH analysed the data from the holistic review of case notes. KH carried out analysis of the infection related data, intensive care mortality data, and hand hygiene related data. JD was responsible for all aspects of the staff and patient surveys. RJL and MDW led on writing of the paper and interpretation of the findings. All authors contributed to the final manuscript. RL is guarantor.

Funding: This study was funded by the Health Foundation and the National Patient Safety Agency. KH was funded by the National Institute for Health Research Collaborations for Leadership in Applied Health Research and Care for Birmingham and Black Country, and AG by the Engineering and Physical Sciences Research Council, Multidisciplinary Assessment of Technology Centre for Healthcare programme. The Centre for Medication Safety and Service Quality is affiliated with the Centre for Patient Safety and Service Quality at Imperial College Healthcare N HS Trust, which is funded by the National Institute of Health Research. The evaluation was sponsored for research governance purposes by the University of Birmingham. The study was designed independently by the researchers. The researchers acted independently but worked collaboratively with the funder. The researchers independently collected, analysed and interpreted the data. The researchers wrote this article independently. The funders were given the opportunity to provide comments before submission. All researchers (apart from CD) had the opportunity to access participant anonymised data.

Competing interest: All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: financial support as specified elsewhere; no financial relationships with commercial entities that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Each substudy had its own ethical approval. The staff and patient surveys were approved by the North West multi-centre research ethics committee and each site granted access to their data. The National Research Ethics Service deemed the case note review as audit/service evaluation and no further ethical approval was required. Permission was also granted from each site to access ICNARC, NOSEC and healthcare associated infection data. Local research governance was followed at each site.

Data sharing: No additional data available, but see full report.[2]

1 Benning A, Ghaleb M, Suokas A, Dixon-Woods M, Dawson J, Barber N, et al. Large scale organisational intervention to improve patient safety in four UK hospitals: mixed method evaluation. *BMJ* 2011;doi:10.1136/bmj.d195.
2 Benning A, Nwulu U, Ghaleb M, Dixon-Woods M, Dawson J, Barber N, et al. A controlled evaluation of the second phase of a complex patient safety intervention implemented in English hospitals. 2010. www.haps.bham.ac.uk/publichealth/psrp/EvalSPI.shtml.
3 Brown C, Hofer T, Johal A, Thomson R, Nicholl J, Franklin BD, et al. An epistemology of patient safety research: a framework for study design and interpretation. Part 2. Study design. *Qual Saf Health Care* 2008;17:163-9.
4 British Orthopaedic Association. Primary total hip replacement: a guide to good practice. 2006. www.boa.ac.uk/site/showpublications.aspx?ID=59.
5 Institute for Healthcare Improvement. How to guide. Prevent surgical site infections. 2007. www.ihi.org/IHI/Topics/PatientSafety/SurgicalSiteInfections/.
6 National Institute for Health and Clinical Excellence. Clinical guideline 65: The management of inadvertent perioperative hypothermia in adults. 2008. www.nice.org.uk/nicemedia/pdf/CG65NICEGuidance.pdf.
7 National Institute for Health and Clinical Excellence. Clinical Guideline 74: Prevention and treatment of surgical site infections. 2010. www.nice.org.uk/nicemedia/pdf/CG74NICEGuideline.pdf.
8 National Patient Safety Agency. Cleanyourhands campaign. 2010. www.npsa.nhs.uk/cleanyourhands.
9 Stone S, Slade R, Fuller C, Charlett A, Cookson B, Teare L, et al. Early communication: does a national campaign to improve hand hygiene in the NHS work? Initial English and Welsh experience from the NOSEC study (National Observational Study to Evaluate the CleanYourHandsCampaign). *J Hosp Infect* 2007;66:293-6.
10 Department of Health. Beds open overnight in England. 2010. www.dh.gov.uk/en/Publicationsandstatistics/Statistics/Performancedataandstatistics/Beds/DH_083781.
11 Harrison DA, Brady AR, Rowan K. Case mix, outcome and length of stay for admissions to adult, general critical care units in England, Wales and Northern Ireland: the Intensive Care National Audit and Research Centre Case Mix Programme Database. *Crit Care* 2004;8:R99-111.
12 Harrison DA, Brady AR, Parry GJ, Carpenter JR, Rowan K. Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. *Crit Care Med* 2006;34:1378-88.

13 Harrison DA, Parry GJ, Carpenter JR, Short A, Rowan K. A new risk prediction model for critical care: the Intensive Care National Audit & Research Centre (ICNARC) model. *Crit Care Med* 2007;35:1091-8.
14 Guzzo RA, Jette D, Katzell RA. The effects of psychologically based intervention programs on worker productivity: a meta-analysis. *Pers Psychol* 1985;38:275-92.
15 West MA, Guthrie JP, Dawson JF, Borrill CS, Carter MR. Reducing patient mortality in hospitals: The role of human resource management. *J Organ Behav* 2006;27:983-1002.
16 Borrill CS, West MA, Shapiro D, Rees A. Team working and effectiveness in health. *Br J Healthcare Manag* 2000;6:364-71.
17 Michie S, West M. Managing people and performance: An evidence-based framework applied to health service organisations. *Int J Manag Rev* 2004;5-6:91-111.
18 Healthcare Commission. Making sense of your staff survey data. 2006. www.cqc.org.uk/usingcareservices/healthcare/nhsstaffsurveys.cfm.
19 Leape LL, Berwick DM. Five years after to err is human: what have we learned? *JAMA* 2005;293:2384-90.
20 Campbell SE, Walke AE, Grimshaw JM, Campbell MK, Lowe GD, Harper D, et al. The prevalence of prophylaxis for deep vein thrombosis in acute hospital trusts. *Int J Qual Health Care* 2001;13:309-16.
21 Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. *N Engl J Med* 1991;324:370-6.
22 Slaughter MJ. Trade liberalization and per capita income convergence: a difference-in-difference analysis. *J Int Econ* 2001;55:203-28.
23 Gallivan S, Taxis K, Dean Franklin B, Barber N. Is the principle of a stable Heinrich ratio a myth? A multimethod analysis. *Drug Saf* 2008;31:637-42.
24 Lilford R, Edwards A, Girling A, Hofer T, Di Tanna GL, Petty J, et al. Inter-rater reliability of case-note audit: a systematic review. *J Health Serv Res Polic* 2007;12:173-80.
25 Lilford RJ, Chilton PJ, Hemming K, Taylor CA, Girling AJ, Barach P. Evaluating policy and service interventions: a framework to guide selection and interpretation of study end points. *BMJ* 2010;341:c4413.
26 Davidoff F. Heterogeneity is not always noise: lessons from improvement. *JAMA* 2009;302:2580-6.
27 Bate P, Mendel P, Robert G. Organizing for quality: the improvement journeys of leading hospitals in Europe and the United States. Radcliffe Publishing, 2008.
28 Department of Health. Saving lives: reducing infection, delivering clean and safe care. 2007. www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_078134.
29 House of Commons Health Committee. Patient safety: sixth report of session 2008-09. Stationery Office, 2009.
30 Dixon-Woods M, Tarrant C, Willars J, Suokas A. How will it work? A qualitative study of strategic stakeholders' accounts of a patient safety initiative. *Qual Saf Health Care* 2010;19:74-8.
31 National Institute for Health and Clinical Excellence. Clinical Guidelines CG50. Acutely ill patients in hospital. Recognition of and response to acute illness in adults in hospital. 2007. www.nice.org.uk/nicemedia/pdf/CG50FullGuidelineShort.pdf.
32 Sacks H, Chalmers TC, Smith H Jr. Randomized versus historical controls for clinical trials. *Am J Med* 1982;72:233-40.
33 Epstein S. The construction of lay expertise: AIDS activism and the forging of credibility in the reform of clinical trials. *Sci Technol Hum Val* 1995;20:408-37.
34 Andriole GL, Crawford D, Grubb RL, Buys SS, Chia D, Church TR, et al. Mortality results from a randomized prostate-cancer screening trial. *N Engl J Med* 2009;360:1310-9.
35 Hillman K, Chen J, Cretikos M, Bellomo R, Brown D, Doig G, et al. Introduction of the medical emergency team (MET) system: a cluster-randomised controlled trial. *Lancet* 2005;365:2091-7.
36 Keogh-Brown MR, Bachmann MO, Shepstone L, Hewitt C, Howe A, Ramsay CR, et al. Contamination in trials of educational interventions. *Health Technol Assess* 2007;11:ix-107.

Accepted: 12 October 2010